

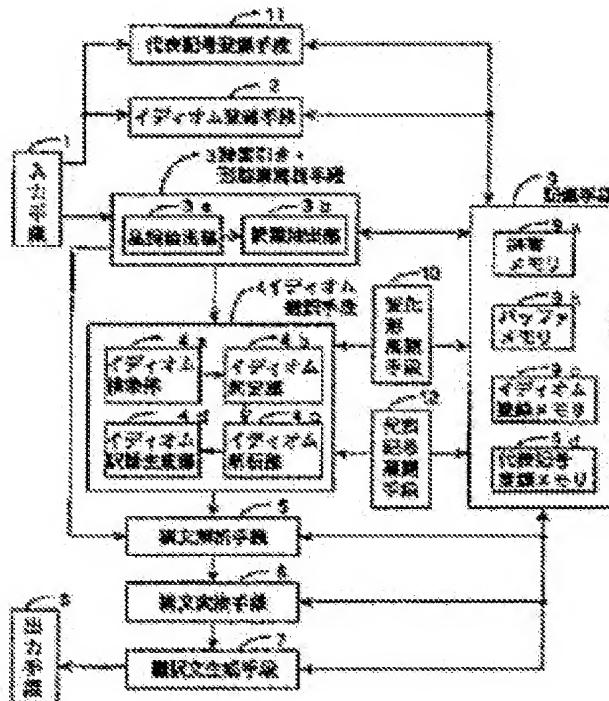
MACHINE TRANSLATION SYSTEM WITH IDIOM REGISTERING FUNCTION

Publication number: JP8055123
Publication date: 1996-02-27
Inventor: OKUNISHI TOSHIYUKI; FUKUMOCHI YOUJI; SADA ICHIKO; KUTSUMI TAKESHI
Applicant: SHARP KK
Classification:
- International: G06F17/27; G06F17/28; G06F17/27; G06F17/28;
(IPC1-7): G06F17/28; G06F17/27
- European:
Application number: JP19940186127 19940808
Priority number(s): JP19940186127 19940808

[Report a data error here](#)

Abstract of JP8055123

PURPOSE: To provide a machine translation system capable of corresponding also to a modification form of a fixed part in an index word of an idiom and leading a user definition mark into a variable part in respect to a machine translation system with an idiom registering function. CONSTITUTION: The machine translation system is provided with an idiom registering means 2 for registering an entry word and a translated word of an idiom so that the fixed part of the idiom is expressed by a normal word, a normal word string or a modification developing mark representing the modified expression of the word or the word string and the variable part of the idiom is expressed by the composite format of a 1st representative mark representing the set of words or word strings sharing a prescribed attribute, a modification form developing means 10 for generating and developing the fixed part of the index word of an idiom to be translated to all previously set modified expressions and an idiom translation means 4 for identifying an input character string or a part of the character string with the index word or the like obtained by developing the fixed part of the index word of the idiom is developed to a modified expression.



Data supplied from the esp@cenet database - Worldwide

【特許請求の範囲】

【請求項1】 文字列および記号を入力する入力手段と、

予め定められた単語又は単語列からなる固定部分と、共通の属性を持つ単語又は単語列に変化可能な可変部分とからなるイディオムに対して、前記固定部分が通常の単語、単語列、又はその単語もしくは単語列の変形表現を代表する変化展開記号によって表現され、かつ前記可変部分が前記の属性を共有する単語又は単語列の集合を代表する第1の代表記号を複数した形式で表現される見出し語とそのイディオムの統語を登録するイディオム登録手段と、

イディオムの登録と翻訳処理に必要な辞書及び処理結果を記憶する記憶手段と、

入力單語列を形態素に分離し、かつ文法解析を行う辞書引き・形態素解析手段と、

翻訳すべきイディオムの見出し語に対してその固定部分を予め設定されたすべての変形表現に生成展開する変化形展開手段と、

入力文字列あるいはその一部分と、登録されたイディオムの見出し語あるいは前記変化形展開手段によってその見出し語の固定部分が変形表現に展開された見出し語との同定を行い、同定されたイディオムの見出し語に対応する文字列の統語を生成するイディオム翻訳手段と、

構文解析手段と、構文変換手段と、翻訳文生成手段と、翻訳文を出力する出力手段とを備えたことを特徴とするイディオム登録機能を有する機械翻訳装置。

【請求項2】 前記変化形展開手段が、イディオムの見出し語の固定部分を、固定部分を構成する単語を活用変化させた表現形式又はその固定部分に動詞もしくは否定動詞を連接させた表現形式に生成展開することを特徴とする請求項1のイディオム登録機能を有する機械翻訳装置。

【請求項3】 新たに定義された属性とその属性値を有する単語又は単語列を代表する第2の代表記号を前記記憶手段に登録する代表記号登録手段と、

翻訳すべきイディオムの見出し語の可変部分に含まれる前記第2の代表記号を定義された属性とその属性値とに生成展開する代表記号展開手段とを備え、

前記イディオム登録手段が、前記入力手段によって入力された第1の代表記号および/または第2の代表記号を用いて表現されるイディオムの見出し語とその統語を登録し、前記イディオム翻訳手段が、入力文字列あるいはその一部分の属性及びその属性値と、前記代表記号展開手段によって生成展開された見出し語の第2の代表記号の属性及びその属性値との同定を行うことを特徴とする請求項1または2記載のイディオム登録機能を有する機械翻訳装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 この発明は、電子化辞書あるいは電子化辞書を搭載した情報検索装置あるいは電子化辞書を搭載した機械翻訳装置に関し、特に、可変部分を含むイディオムを登録し検索・翻訳することのできるイディオム登録機能を有する辞書検索装置に関する。

【0002】

【従来の技術】 現在実用化されている言語処理装置には、人間の文書作成活動を支援するためのワードプロセッサや、ある言語で書かれた文書を別の言語に翻訳するための機械翻訳装置などがある。これらの言語処理装置には、それぞれの目的に応じた情報を含めた辞書が備えられている。ここでいう辞書とは、見出し語とそれに付帯する各種の情報の組とを1単位の項目としたものを多数統合し、見出し語を用いて所取の項目を容易に検索できるように系統立てて並べたものである。

【0003】 辞書は、原則として機械可読な不揮発性の媒体に機械可読な形式で記録される。このような辞書を、電子化辞書と呼ぶことにする。電子化辞書を機械翻訳において用いる場合には、見出し語としては原語（ソース原語）の単語列（1単語のみのものも含む）が用いられ、その単語列に付帯する各種の情報として、見出し語の品詞、形態属性、説明、説語の品詞等の情報が用いられる。

【0004】 このような言語処理装置を用いて利用者が処理あるいは作成しようとしている文書に、この装置に備えられた辞書に見出し語として記載されていない単語が含まれている場合には、作業効率が著しく低下してしまう。そのため、辞書に収録する見出し語は、より多いほうが好ましい。また、機械翻訳の場合には、原語の各単語のみではなく、イディオムを見出し語として採用し、対応するターゲット言語の言い回し等をペアとして、このようなペアをできるだけ多数登録しておくことが翻訳効率の上では望ましい。

【0005】 通常イディオムには、数詞、所有格代名詞、再帰代名詞など、主語や他の語との関係においてその形を変えるイディオムが多い。例えば〔do one's best〕中のone'sは主語に応じてyour, my, his, herなどの所有格代名詞となる。翻訳処理及びに辞書開発の効率上、このようなイディオムは、具体的な語を入れたイディオムを全て列举するのではなく、見出し語の一部分の単語としてある文法特徴を共有する単語や句ならば任意の単語が入るような形で登録できるのが好ましい。以下では、ある文法特徴を共有する単語や句ならば任意の単語が入る部分を可変部分と呼び、それ以外のイディオムの骨格となる単語や単語列の部分を固定部分と呼ぶ。

【0006】 また、1つのイディオムには、複数の可変部分が存在するものがあり、可変部分には単語だけでなく名詞句や文が適用される場合もある。このような種々の可変部分を表現するために、*に続く記号（以下この記号を代表記号と呼ぶ）を導入する機械翻訳装置が提案

3

されている。

【0007】たとえば「～より大き～」という日本語訳をもつ英単語イディオム“N times-as-as~”を登録する場合、次のように記述することができる。

英単語 [*3 times as *ad as *CN]

訳語 *3 より *ad 大きい

見出し(英単語列)の中で、先頭に“*”のついた単語(*3, *ad, *CN)、すなわち代表記号が可変部分であり、それ以外の単語(times, as)が固定部分である。

【0008】可変部分では、代表記号で表す品詞(上例では、n:数詞、ad:形容詞、c:文、CN:名詞句、CN:cまたは3)の任意の単語と適合できるのに対し、固定部分ではその表記を持つ単語とでないと適合できない。なお、可変部分に指定できる代表記号は予めシステムに定義されたものである。
*

[This apple is three times as big as that orange.]

*3 *times *1 *ad *1 *CN

ここで、可変部分の“*3”がマッチングし、“*ad”と“*big”がマッチングし、“*CN”と“*that orange”がマッチングする。また、“*times as”と“as”がイディオムの固定部分(※記持部)としてマッチングする。

【0011】次に、イディオムの可変部分の訳を生成する。
*

「この りんごは あのオレンジ より 3 倍 大きい」
*3 *times *1 *ad *1 *CN

【0012】

【発明が解決しようとする課題】しかしながら、以上に示した機械翻訳システムでは、

(1) イディオム見出しの固定部分に、活用変化が繰り返さない。

(2) イディオム見出しの可変部分に、システムで定義された代表記号以外の記号が使えない。

という2つの制限がある。このため、イディオムに登録する見出し語が増大し、その結果記憶容量及び検索時間が拡大するという問題や、利用者、すなわち辞書にイディオムを登録する者の負担が大きくなるという問題が発生する。

【0013】固定部分と可変部分から成るイディオム見出し語の固定部分と入力文のマッチングは文字列だけで比較するので、單複数変化する名詞、活用変化する動詞、助動詞、形容詞、また、助動詞や否定副詞(not)が付加する単語を含む入力文の場合には、通常よく使われる基本形のイディオムを登録するだけではマッチングが失敗することとなる。そのため、以上に示したような変化形を全て異なる見出しとして列挙する必要があった。

【0014】『～することはできない』という日本語訳を持つ英単語イディオム“There is no ~ing.”を例に説明する。このイディオムは、次のような形式で登録される。

英単語 [There is no *Ving]

* [0009] 今、

[This apple is three times as big as that orange.
]

という英文が入力されたとし、以下に機械翻訳処理の概要を以下に示す。まず、各単語の絆書きが行われる。その結果、次のような単語情報が得られる。

three 数詞(n)

big 形容詞(ad)

that 冠詞(d)

orange 名詞(n)

【0010】次に、イディオムの検索と解釈処理が実行され、固定部分ならびに可変部分のマッチングが行われる。このとき、入力文の一部が上記のように登録されたイディオムに適合することがわかる。

[This apple is three times as big as that orange.]

*3 *times *1 *ad *1 *CN

*3 *CN あのオレンジ

*3 3

*3 *ad 大きい

最後に、固定部分の訳と合成することで入力文に対する日本語文を生成する。

「この りんごは あのオレンジ より 3 倍 大きい」
*3 *times *1 *ad *1 *CN

訳語 *Ving ことができない。

ここで、“*Ving”は「動詞の進行形」を表す代表記号である。

【0015】しかし、“There was no going to school yesterday.”という文はこのイディオムを含むにも関わらず過去形の文であるため、文中の“There was no”が見出しの固定部分の“There is no”と一致できない。すなわち、一部に動詞等の語形変化する単語を含んだイディオムの場合、変化形を全て展開して登録する必要があり、この例では[There was no]や[There will be no]という見出しのイディオムを別に登録する必要があった。

【0016】固定部分と可変部分から成るイディオム見出しの可変部分は、あらかじめ機械翻訳システムで定義されたもののみが使用される。これは、幅広い利用者がターゲットである商用化システムにおいて、利用者のあらゆる要求を予測し代表記号を全て準備しておくのは現実には不可能であるし、非常に特殊なものまでシステムで定義しておくのも効率上問題があったからである。

【0017】このような従来の機械翻訳システムにおいて、所望する代表記号が定義されていなかった場合について、「できる限り～」という日本語訳を持つ英単語イディオム“as～as～can”を例に説明する。3つのasの間の「～」に入れることができる語句は、形容詞原級か、形容詞原級で標識される名詞句であるが、もし、シ

5

システムに定義された代表記号に「形容詞を含む名詞句」という記号がなかったとすると、利用者はシステムが定義している代表記号で最も近いもの、例えば、次のように、*3(名詞句)とだけ指定しているイディオムを使用するしかない。

英単語 [as .as as .as can]

翻訳 できる限り *

ここで、*3は主格代名詞を表す代表記号である。

【0018】しかし、このような近似的な翻訳では実際 *

[I regard the person as policeman as you can see.]

*3 細 犯 種

というように、policeman が粗とマッチしてしまい、この文がイディオムを含んでいると誤って認識してしまう。

【0019】さらに、イディオムと解釈したas policeman as as you can が本語でsee が動詞からなる関係代名詞の先行詞がthe personであるという誤った解析がたまたま成立し、結果的に次のような翻訳を生成してしまう。

「私はできる限り警察を見る人を考慮する。」

【0020】このようにシステムで定義された代表記号しか使えないため、近似的な指定しかできず、利用者が意図しなかった誤訳が生じる場合もある。また、指定を詳細にすると見出しが複雑になる場合が多い。このことは、無理した詳細な条件が複数現れた場合に特に顕著になる。このように、従来の機械翻訳装置では、以上述べたような2つの制限があるため、イディオムに登録する見出し語が増大しその結果記憶容量及び検索時間が増大するという問題や、利用者の負担が大きくなるという問題が発生していた。

【0021】そこで、この発明は、以上のような事情を考慮してなされたものであり、イディオムの見出し語の固定部分の変形形にも対応でき、可変部分に利用者定義の記号を導入することができるイディオム登録機能を持つ機械翻訳装置を提供することを目的とする。

【0022】

【課題を解決するための手段】図1に、この発明の基本構成ブロック図を示す。同図において、この発明は文字列および記号を入力する入力手段1と、予め定められた単語又は単語列からなる固定部分と、共通の属性を持つ単語又は単語列に変形可能な可変部分とからなるイディオムに対して、前記固定部分が通常の単語、単語列、又はその単語もしくは単語列の変形表現を代表する変形展開記号によって表現され、かつ前記可変部分が既定の属性を共有する単語又は単語列の集合を代表する第1の代表記号を複合した形式で表現される見出し語とそのイディオムの訳語を登録するイディオム登録手段2と、イディオムの登録と翻訳処理に必要な辞書及び処理結果を記憶する記憶手段3と、入力単語列を形態素に分解し、かつ文法解析を行う終点引き・形態素解析手段3と、翻訳すべきイディオムの見出し語に対してその固定部分を予

6

*にはイディオムを含んでいない文でも誤って認識して、翻訳精度の低下を招く場合がある。例えば、

[I regard the person as policemen as you can see.]

あなたがわかるので、私はその人を警官とみなす。

という入力文は、2つのasの間の名詞句が形容詞を含んでいないため本来ならば上記イディオムの文ではないが、上記登録では

[I regard the person as policeman as you can see.]

*3 細 犯 種

め設定されたすべての変形表現に生成展開する変形展開手段10と、入力文字列あるいはその一部分と、登録されたイディオムの見出し語あるいは前記変形展開手段10によってその見出し語の固定部分が変形表現に展開された見出し語との同定を行い、同定されたイディオムの見出し語に対応する文字列の訳語を生成するイディオム翻訳手段4と、構文解析手段5と、構文変換手段6と、翻訳文生成手段7と、翻訳文を出力する出力手段8とを備えたことを特徴とするイディオム登録機能を有する機械翻訳装置を提供するものである。

【0023】また、前記変形展開手段10が、イディオムの見出し語の固定部分を、固定部分を構成する単語を活用変化させた表現形式又はその固定部分に助動詞もしくは否定副詞を追加させた表現形式に生成展開するようにもよい。

【0024】また、新たに定義された属性とその属性値を有する単語又は単語列を代表する第2の代表記号を前記記憶手段9に登録する代表記号登録手段11と、翻訳すべきイディオムの見出し語の可変部分に含まれる前記第2の代表記号を登録された属性とその属性値とに生成展開する代表記号展開手段12とを備え、前記イディオム登録手段2が、前記入力手段1によって入力された第1の代表記号および/または第2の代表記号を用いて表現されるイディオムの見出し語とその訳語を登録し、前記イディオム翻訳手段4が、入力文字列あるいはその一部分の属性及びその属性値と、前記代表記号展開手段12によって生成展開された見出し語の第2の代表記号の属性及びその属性値との同定を行なうようにもよい。

【0025】また、前記記憶手段9は、入力された文字列の翻訳を行うための文法および語彙情報を持つ辞書メモリ9aと、訳語生成に至るまでの処理の結果を記憶するバッファメモリ9bと、前記イディオム登録手段2によって登録されたイディオムを記憶するイディオム登録メモリ9cとから構成することが好ましい。

【0026】前記経路引き・形態素解析手段3は、入力された文字列を単語に分解し各単語の品詞情報を生成する品詞抽出部3aと、各単語の訳語の接続の接続を生成する訳語抽出部3bとから構成することが好ましい。

【0027】前記イディオム翻訳手段4は、前記イディ

オム登録メモリを検索し分解された単語列と表現形式が一致可能なイディオムの見出し語の候補を選択するイディオム検索部4 aと、イディオムの中の代表記号の位置に相当する単語又は単語列の属性が、代表記号に与えられた属性に一致するイディオムの見出し語をイディオム候補の中から一つに特定するイディオム同定部4 bと、代表記号に対応する単語又は単語列の構文を解析してイディオム全体の文構成を生成するイディオム解析部4 cと、イディオムの文構成を基に、入力された単語列のイディオム部の訳語を生成するイディオム訳語生成部4 dとから構成することが好ましい。

【0028】ここで、図1において入力手段1としては、キーボード、又はマウス、ペンあるいはトラックボールなどのポイントティングデバイスが用いられるがこれに限定されるものではなく、その他の入力装置を用いてもよい。また、記憶手段9は通常ROM、RAM、フロッピーディスク等はハードディスク等が用いられるが、これに限定されるものではなく、その他の記憶装置を用いてもよい。特に、静音メモリ9 aはRAMが好ましく、パッファメモリ9 b、イディオム登録メモリ9 c及び代表記憶メモリ9 dはRAMが好ましい。

【0029】また、イディオム登録手段2、代表記号登録手段11、辞書引き・形態素解析手段3、イディオム翻訳手段4、構文解析手段5、構文変換手段6及び翻訳文生成手段7としては、通常CPUが用いられ、ROM、RAM、I/Oインターフェイス等の周辺回路を含んだマイクロコンピュータを用い、ROM又はRAMにはこの文書処理装置の動作を制御するプログラムが内蔵されていることが好ましい。

【0030】
【作用】イディオムの見出し語および訳語を登録する場合、イディオムの固定部分は通常の単語、単語列又はその単語もしくは単語別の変形表現を代表する変化展開記号によって表現され、イディオムの可変部分は所定の属性を共有する単語又は単語列の集合を代表する第1の代表記号を複合した形式で表現されて、イディオム登録手段2が記憶手段9に登録する。

【0031】これにより、可変部分の属性が共通し、さらに固定部分が種々の変形表現されることのある複数箇のイディオムの見出し語を1つの見出し語で登録することができる。すなわち、この発明によれば、イディオムの見出し語としてその固定部分及び可変部分に対して考えられるあらゆるパターンを登録する必要はなく、登録されるイディオムの見出し語の増大を抑えることができる。

【0032】入力された文字列を単語に分解し、分解された単語別の一の表現形式と一致可能なイディオムの見出し語の候補を検索し、さらにその見出し語の固定部分の中に変化展開記号がある場合には、変化展開手段によって生成展開された変形表現とその変化展開記号が

存在する位置に相当する単語又は単語列との同定を行い、イディオムを特定する。

【0033】以上のように、この発明によれば、入力文のうちあるイディオムの固定部分に相当する単語又は単語列が、そのイディオムの見出し語として登録されている単語又は単語列とは完全に一致しないがその変形表現と一致する場合にも、イディオムの同定をすることができる。

【0034】また、イディオムの可変部分は新たに定義された属性とその属性値を有する単語又は単語列を代表する第2の代表記号を含む形式で表現されて記憶手段9に登録される。

【0035】そして、入力された文字列を単語に分解し、分解された単語別の一の表現形式と一致可能なイディオムの見出し語の候補を検索し、さらにその見出し語の可変部分の中に第2の代表記号がある場合には、代表記号展開手段によって生成展開された属性及びその属性値と、その第2の代表記号が存在する位置に相当する単語又は単語列の属性及びその属性値との同定を行い、イディオムを特定する。

【0036】以上のように、この発明によれば、新たに定義された属性及びその属性値を有する単語又は単語列を代表する第2の代表記号を用いてイディオムの見出し語の可変部分を表現し、かつイディオムの同定を行うので、イディオムとして登録する見出し語の長さを抑えることができると共に、利用者自身にとって必要な、あるいは、ある分野の文章に特有な表現形式を持つイディオムの登録及び同定をすることができる。

【0037】

【実施例】以下、図に示す実施例に基づいて、この発明を詳述する。なお、これによってこの発明が設定されるものではない。実施例の説明の前に、機械翻訳の概念について簡単に説明する。図2を参照して、機械翻訳において行なわれる解析処理には、様々な解析レベルがある。機械翻訳は、図2の左上に示されるソース言語が入力された場合に、各レベルの処理を順に行なって最終的に図2の右側に示されるターゲット言語を得るための処理である。すなわちソース言語が入力されると、まずレベルL1の辞書引き処理、レベルL2の形態素解析処理、レベルL3の構文解析処理、…と処理が進められ、最終的にレベルL10の形態素生成処理が行なわれてターゲット言語が生成される。

【0038】機械翻訳は、どのレベルの解析処理まで行なうかによって、大きく次の2つに分けられる。第1は、レベルL6に示されるソース言語およびターゲット言語のどちらにも依存しない概念である中間言語で解析し、そこからレベルL7の文脈生成、レベルL8の意味生成、レベルL9の構文生成、レベルL10の形態素生成へと進み、ターゲット言語を生成していくピボット方式である。第2は、上述のレベルL2の形態素解析、

9.

レベル L 3 の構文解析、レベル L 4 の意味解析およびレベル L 5 の文脈解析のいずれかまで解析を行なってソース言語の内部構造を得、次に、得られたソース言語の内部構造と同じレベルのターゲット言語の内部構造に変換した後、ターゲット言語を生成するトランスマスター方式である。

【0039】以下、図 2 に示される各解析処理の内容について説明する。

(1) 辞書引き・形態素解析

ここでは、形態素が格納された辞書を参照しながら入力された文章を形態素列（単語列）に分割し、この各単語に対する品詞などの文法情報をおよび統系を得、さらに時制・人称・数などを解析する処理が行なわれる。

【0040】(2) 構文解析

ここでは、単語間の係り受けなどの文章の構造（構造解析木）を決定する処理が行なわれる。

(3) 意味解析

複数の構造解析の結果から、意味的に正しいものとそうでないものとを判別する処理が行なわれる。

(4) 文脈解析

文脈解析処理では、入力された文章の話題を理解し、入力文箇中に含まれる省略部分や意味などを取扱う処理が行なわれる。

【0041】次に、図 3 に示すこの発明の一実施例である機械翻訳装置のブロック圖について説明する。同様において、3-1 はメイン CPU (中央處理装置)、3-2 はメインメモリ、3-3 は CRT (陰極線管) や LCD (液晶表示装置) などからなる表示装置、3-4 はキーボード、3-5 は翻訳モジュール、3-6 は翻訳モジュール 3-5 に接続された翻訳用の辞書、文法規則および本変換構造規則などを格納している辞書メモリ、3-7 は上記構成部品を接続するバスである。

【0042】また、辞書メモリ 3-6 には、イディオムや、利用者が独自に定義した代表記号を格納しておくことのできる記憶領域を備える。CPU 3-1 は、イディオムの登録及び代表記号の登録の処理と、後述する翻訳モジュール 3-5 の処理の制御を行う。

【0043】翻訳モジュール 3-5 は、ソース言語の文章が入力されると、それを所定の手順で翻訳してターゲット言語を出力するものである。すなわち、キーボード 3-4 から入力されたソース言語はメイン CPU 3-1 の制御により翻訳モジュール 3-5 に送られる。翻訳モジュール 3-5 は辞書メモリ 3-6 に記憶されている辞書、文法規則および本構造変換規則等を用いて、入力されたソース言語を後に詳述するようにしてターゲット言語に翻訳する。その結果は、メインメモリ 3-2 に一旦記憶されると共に、表示装置 3-3 に表示される。

【0044】図 4 に翻訳モジュール 3-5 のブロック圖を示す。翻訳モジュール 3-5 は、バス 3-7 に接続され、バス 3-7 を介して入力されるソース言語を、所定の翻訳ア

ログラムに従って翻訳してターゲット言語としてバス 3-7 に出力するための翻訳 CPU 4-6 と、バス 3-7 に接続され、翻訳 CPU 4-6 で実行される翻訳プログラムを格納するための翻訳プログラムメモリ 4-6 と、入力されたソース言語の原文を各単語ごとに格納するためのパッファ A (4-0) と、パッファ A (4-0) に格納された各単語につき、辞書メモリ 3-6 に含まれる辞書を参照して得た各単語の品詞、説明などの情報を格納するためのパッファ B (4-1) と、ソース言語の構造解析木に関する情報を格納するためのパッファ C (4-2) と、ソース言語の構造解析木から変換されたターゲット言語の構造解析木を格納するためのパッファ D (4-3) と、パッファ D (4-3) に格納されたターゲット言語の構造解析木に適切な助詞語（日本語ならば助詞や助動詞など）を補充して、ターゲット言語の形として整えられた文章を格納するためのパッファ E (4-4) とを含む。

【0045】以上のような構成を持つ翻訳モジュール 3-5において、少なくとも图 2 に示したレベル L 3 の構文解析のレベルまでの解析を行うものとする。ここで、翻訳処理手順を記述した並記翻訳プログラムは、辞書引き・形態素解析部、イディオム翻訳部、構文解析部、構文変換部、翻訳文生成部、変化形接頭部及び代表記号接頭部から構成される。

【0046】以下、図 3～図 10 を参照して、本実施例の機械翻訳装置による英日翻訳の動作を説明する。ここでは、イディオムを含まない英文 "This is a pen." を例にとって、この英文を日本文に翻訳する動作の概要を示す。

【0047】まず、読み込まれた原文は形態素解析によって形態素に分解され、図 5 に示されるようにパッファ A (4-0) (図 4 参照) に格納される。続いて翻訳プログラムメモリ 4-6 に記憶されたプログラムに基づく翻訳 CPU 4-6 の制御の下に、辞書引き・形態素解析部によつて、パッファ A (4-0) に格納された原文の各単語ごとに、辞書メモリ 3-6 に格納されている辞書を参照することにより各単語の品詞などの情報を得られる。たとえば、その情報の一部である品詞情報は、図 6 のようにパッファ B (4-1) に格納される。

【0048】ここで、"this" の多品詞語であって代名詞、指示形容詞の 2 つの品詞を持つ。また "is" の品詞は動詞である。同様に "a"、"pen" についてもそれぞれの品詞がパッファ B (4-1) に格納される。"this" は多品詞語であるが、文中の品詞が何であるかについては、翻訳プログラムのうち構文解析部に相当する処理によって一意に決定される。

【0049】翻訳プログラムのうち構文解析部に相当する処理においては、辞書メモリ 3-6 に格納された辞書および文法規則に従って、各単語間の係り受け関係を示す構造解析木がたとえば図 7 に示されるように決定される。この構文解析結果は図 4 のパッファ C (4-2) に格

11

納される。

【0050】構造解析木の決定は次のようにして行なわれる。辞書メモリ36に格納された文法規則から、英語に関する文法規則として次のようなものが得られる。

文→主部、述部

主部→名詞句

述部→動詞、名詞句

名詞句→代名詞

名詞句→冠詞、名詞

【0051】この規則のうちたとえば1つ目の規則は、「文は主部と述部からできている。」ということを表わす。以下、これらの規則に従って構造解析木が決定される。なお、このような文法規則は同じように日本語についても用意されており、英語の文法規則と日本語の文法規則との間で対応づけがなされている。

【0052】翻訳プログラムのうち、構文変換部に相当する処理においては、辞書メモリ36の木構造変換規則を用いて、入力された英文の構造解析木(図7参照)の構造が、図8に示される日本文に対する構文解析木の構造に変換される。得られた結果は図9に示されるバッファD(43)に格納される。この説明において用いられている例文 "This is a pen." は、この変換によって日本語文字列「これ ペン である」に変換されたことになる。

【0053】翻訳プログラムのうち翻訳文生成部に相当する処理を行なう部分は、得られた日本語文字列「これ ペン である」に適切な助詞「は」や助動詞をつけることにより、図9に示されるような日本語の形にし、図9のバッファE(44)に格納する。この得られた日本文「これはペンである。」は、図3に示される翻訳モジュール35から出力され、メインメモリ32に格納されるとともに、表示装置33に表示される。

【0054】以上が、イディオムを含まない文の翻訳処理の概要であるが、イディオムを含む文の翻訳処理においては、上記処理のほか、イディオム翻訳部におけるイディオムの同定、解析及び翻訳の生成処理が行われ、さらに、イディオム翻訳部の処理に関連して、変化形展開部及び代表記号展開部の処理が行われる。

【0055】ここで変化形展開部は、後述するように、イディオムの固定部分に対して活用変化などの変化形を考慮したマッチング処理を行うものである。また、代表記号展開部は、利用者が自由に定義した代表記号に対してマッチング処理を行うものである。以上の各部の処理は、翻訳モジュール35の翻訳CPU45によって翻訳プログラムの手順に従って実行される。

【0056】次に、図10～14を用いて、見出し語のうち固定部分を変化形展開記号で表現したイディオムの登録について説明する。変化形展開記号は、次のような記号である。

*品詞記号 (單語)

12

なお、「品詞記号」は、変化対象の「単語」が多品詞だった場合に、どの品詞で変化されるかを指定するためのものである。

【0057】たとえば、"as~as can be" というイディオムは次のように記述される。

英単語 [as * as as * (can) be.]

翻訳 この上なく*で

*は助動詞を表す品詞記号であり、助動詞としてのcanの過去形couldであっても、このイディオムであることを表している。品詞記号*を指定するのは、canが多品詞だからである。

【0058】すなわち、canは名詞又は動詞としての用法もあり、品詞の指定がないと、名詞として変化(cansなど)させるべきか、又は動詞として変化(caned)させるべきかを特定するのが困難だからである。

【0059】また、ここで最後のbeには変化展開指定がない。これは、canまたはcouldのどちらかであろうが、be動詞は原形しかありえないからである。このように、変化形を持つが見出し以外の形は認識してはいけない時には、変化展開指定をしない。なお変化させるイディオム登録の方が多い場合には、変化指定のデフォルトを邊にして、無変化の場合に指定させてよい。

【0060】以下に、

[I was as happy as could be.]

という入力文があった場合を例にとり、このイディオムを使った翻訳内容を説明する。図10に、イディオム部分の翻訳処理のフローチャートを示す。

【0061】まず、図10のステップS111～S112において、入力文の先頭単語から順次辞書引きが行われる。代表記号を含んだイディオムも他の単語と同様に基本辞書に登録されているので、3単語目(s=3の時)asの辞書引き中に、[as *as * (can) be] の見出しが検索される(ステップS113、S117、S118)。

【0062】次に、ステップS114において、イディオムの見出し中の各単語と入力文の単語の間でマッチングが行われる。イディオム中の固定部分の単語と入力文の単語の間のマッチングは文字列比較だけで高速に処理できるので、最初に固定部分だけのマッチング処理を行う。

【0063】ここで、この例では、

[as *as as * (can) be.] と "as happy as could be."

の間での固定部分のマッチング処理に入る。

【0064】図11に、この固定部分のマッチング処理のフローチャートを示す。まず、イディオム中の単語番号を示す変数pを初期化する(ステップS121)。ステップS122において、イディオム中の単語番号pの単語W_pが固定部分であるかどうか判断し、固定部分でない、すなわち可変部分である場合は、ステップS13

13

2. S 1.3.2へ処理を進め、すべての単語が調べられるまで、処理を繰り返す。

【0065】また、単語W₁が確定部分である場合は、ステップS 1.2.3へ進み、単語W₁が変化形展開指定かどうか判断する。単語W₁が変化形展開指定の場合は、ステップS 1.2.5、S 1.2.6へ進み、活用変化させたマッチング処理に入るが、単語W₁が変化形展開指定でない場合は、単語W₁とW₂との比較を行い(ステップS 1.2.4)、一致する場合は、次の単語に対するマッチング処理を繰り返す(ステップS 1.3.6～1.3.3、S 1.2.2)。

【0066】上記のイディオムの場合、イディオム見出し先頭の“as”と入力単語が一致しているので、イディオム見出しの次の単語“*a”に処理が移る。(ステップS 1.2.2～S 1.2.4、S 1.3.0～S 1.3.3)。“*a”は可変部分を表す代表記号なので、さらに次の単語“as”に処理が移る(ステップS 1.2.3、S 1.3.3、S 1.3.3)。ここで、“as”は変化形展開指定を含まないので文字列比較を行い(ステップS 1.2.4)、一致していることがわかる。

【0067】次に、イディオムと入力文のそれぞれ次の単語、“as (can)”と“could”的マッチングに処理が移る(ステップS 1.3.0～S 1.3.3)。“as (can)”は変化形展開指定なのでステップS 1.2.6へ進み、文字列の比較に入る前に、“as”の助動詞としての活用変化を行う。

【0068】活用変化は時制や单複などの語尾変化の他に、助動詞や“be”的付加も考慮するので、形態素解析が持つ語尾処理テーブルだけでなく、図1.3に示す変化形テーブルを使って変化させる。本例では、“can”を変化形展開した“could”“can”に入力の“can”が含まれることがわかる(ステップS 1.3.6)ので、次の単語のマッチングを調べていく(ステップS 1.3.0～S 1.3.3)。

【0069】イディオムと入力文のそれぞれ、次の入力単語(“be”)とbeのマッチングに処理が移る。これも同様に一致していることがわかり、結局、確定部分のマッチングは成功することがわかる(ステップS 1.3.3において成功終了)。

【0070】次に、図1.0の可変部分のマッチング処理(ステップS 1.1.5)に入る。図1.2に、この可変部分のマッチング処理のフローチャートを示す。最初に、可変部分の辞書引きを行う(ステップS 1.4.1)。確定部分のマッチングの際に、代表記号*aの対象単語が確定部分に検まれたhappyでありえることがわかっているので、happyの辞書引きを行なう。

【0071】次に、代表記号の中にユーザ代表記号があるかどうか調べ(ステップS 1.4.2)。もしであればユーザ代表記号を定義本体部に置換する(ステップS 1.4.3)。次に、形態素レベルでチェックできるかどうかを

14

調べる(ステップS 1.4.4)。“*a”はシステム定義の単語品詞なので、“happy”的品詞が形容詞であることを確認し(ステップS 1.4.5)、可変部分のマッチングが終了する。

【0072】もし可変部分の対象単語が複数の単語からなる場合、すなわち代表記号が句詞性の場合には、構文処理が呼ばれ、肯定の属性がチェックされる(ステップS 1.4.6)。以上で、入力文がイディオム

英単語 [as *a as *x (can) be.]

翻訳 この上なく幸福で

を含むことがわかったので、最後に、次のようなイディオム部分の訳文を生成し、訳パッファに格納しておき、イディオム処理が完了する(ステップS 1.1.6)。

イディオム部分 [as happy as could be]

訳文 この上なく幸福で

【0073】さらに処理を繰り返し(ステップS 1.1.7、S 1.1.8)。イディオムの次の単語から辞書引きを再開する。この例はイディオムの範囲が文末までなのでこの時点で辞書引きが完了する。以降、イディオム以外の“I was”的単語列に関して、通常の構文解析、構文変換が行われ、最後に、翻訳文生成処理でイディオム以外の日本語訳「私は～あった」とイディオム部分の日本語訳「この上なく幸福で」の合成が行なわれ、次のような文全体の訳文が得られる。

「私は、この上なく幸福であった。」

【0074】次に、利用者が登録した代表記号を用いたイディオムの例と、そのイディオムを用いた文の翻訳処理について説明する。利用者が新しい代表記号を登録するために、次のような書式を利用するものとする。

③ (新記号) “:” (定義本体部)

ここでこのような書式で記述された代表記号は、CPU 3.1によって辞書メモリ 3.6に格納される。

【0075】また、定義本体部は次のように記述する。

“:” (文法属性) “/” (値) “,” (文法属性)

“/” (値) “,” (文法属性)

指定できる文法属性と値(属性値)として、利用者には、翻訳システムが内部的に定めているあらゆるバラメータを開放する。これにより、文法や意味の詳細な制約を使って代表記号を定義でき、簡単にイディオム登録に利用できるようになる。属性と値は、例えば、図1.4のような属性を考えることができる。

【0076】このように形態素解析以外の種々のレベルの属性と値も利用者が指定できるようになると、辞書引きの段階だけでは、ある単語列が利用者が定義した代表記号かどうかのチェックができなくなる。例えば、ある単語列の品詞が「名詞句」であるかどうかは構文解析の段階まで進まないとわからない。図1.4の各属性の値にはそれぞれどの段階でチェックできるかが明記されている。

④ 【0077】以下では、

15

"as 形容詞 a 異詞無の単数名詞句 as ... can"
「この上なく～」

というイディオムに基づき説明する。このイディオムの語彙は特殊であるため、「冠詞無の単数名詞句」という文法的制約の指定が必要になる。

【0078】以下に、この文法的制約を表す代表記号の登録処理と翻訳処理について説明を行なう。図14の表を用いると、品詞(cat)が名詞句(n)で、活用形(Inf)が単数(sg)、用法(use)が冠詞なし(detail)という「冠詞無の単数名詞句」を表す代表記号は、
`英語 ::= [cat/n, Inf/sg, use/detail]`

と登録できる。

【0079】この代表記号を用いることで、上記イディオムは次のように簡単に登録できる。

英単語 [as *a a *sg as *can] be]

説文 この上なく*a *sg

【0080】次に、このイディオムを用いた翻訳処理を図12を用いて説明する。

`[I bought as large a hat as could be.]`

があったとすると、まずイディオムの検索から、固定部分のマッチングまでの処理が、上記実施例と同様に行なわれる。

【0081】次に、可変部分のマッチングに移る。例では、このイディオムの中に入利用者が定義した代表記号を含んでいるので（ステップS142），利用者代表記号を定義本体部に置換する処理を行なう（ステップS143）。

【0082】本例の代表記号*sgの定義本体部の場合には、

`[cat/n, Inf/sg, use/detail]`

のように展開される。図14を参照することで、このうち、cat/n, Inf/sg は形態素解析レベルでチェックできる（ステップS145）のに対して、use/detailは構文解析まで進んで初めて、冠詞無名詞であることがわかる（ステップS146）。すなわち本例では、“a hat”的“a”はイディオム見出し中に含まれるので、可変部分は“hat”だけになり、冠詞無名詞(*sg)であることわかる。

【0083】上記のようにマッチング処理が成功し、イディオムを認識した後は、上記実施例と同様に処理が進み、イディオム部分の訳「この上なく大きな帽子」が得られる。

イディオム部分 [as large a hat as could be]

説文 この上なく大きな帽子

さらに、文全体の説文

「私はこの上なく大きな帽子を買った。」

が得られる。

【0084】

【発明の効果】この発明によれば、イディオムの見出し語の固定部分を変化展開記号を含む表現で登録し、かつ

16

変化展開記号で表現された固定部分を変形表現に生成展開し、この変形表現と変化展開記号が存在する位置に相当する単語又は単語列との同定を行うため、固定部分が種々の変形表現されることのある複数個のイディオムの見出し語を、1つの見出し語で登録することができる。よって、固定部分の種々の変形表現について考えられるあらゆるパターンを登録する必要なく、登録されるイディオムの見出し語の拡大を押さえることができ、さらにイディオム記憶容量及び検索時間の抑制である。

【0085】また、この発明によれば、新たに定義された属性及びその属性値を有する単語又は単語列を代表する第2の代表記号を用いてイディオムの見出し語の対変部分を表現し、かつイディオムの同定を行うので、イディオムとして登録する見出し語の長さを抑えることができると共に、利用者自身にとって必要な、あるいは、ある分野の文章に特有な表現形式を持つイディオムの登録及び同定をすることができる。

【図面の簡単な説明】

【図1】この発明の機械翻訳装置の基本構成を示すブロック図である。

【図2】機械翻訳の概念を模式的に示す図である。

【図3】この発明の一実施例の構成ブロック図である。

【図4】図3に示される翻訳モジュール3.5のブロック図である。

【図5】パッファAの格納内容を模式的に示す図である。

【図6】パッファBの格納内容を模式的に示す図である。

【図7】パッファCの格納内容を模式的に示す図である。

【図8】パッファDの格納内容を模式的に示す図である。

【図9】パッファEの格納内容を模式的に示す図である。

【図10】辞書引き・イディオム処理を示すフローチャートである。

【図11】固定部分のマッチング処理を示すフローチャートである。

【図12】可変部分のマッチング処理を示すフローチャートである。

【図13】変化形テーブルを模式的に示す図である。

【図14】代表記号として指定できる属性を模式的に示す図である。

【符号の説明】

1 入力手段

2 イディオム登録手段

3 辞書引き・形態素解析手段

3 a 品詞抽出部

3 b 例語抽出部

4 イディオム翻訳手段

17

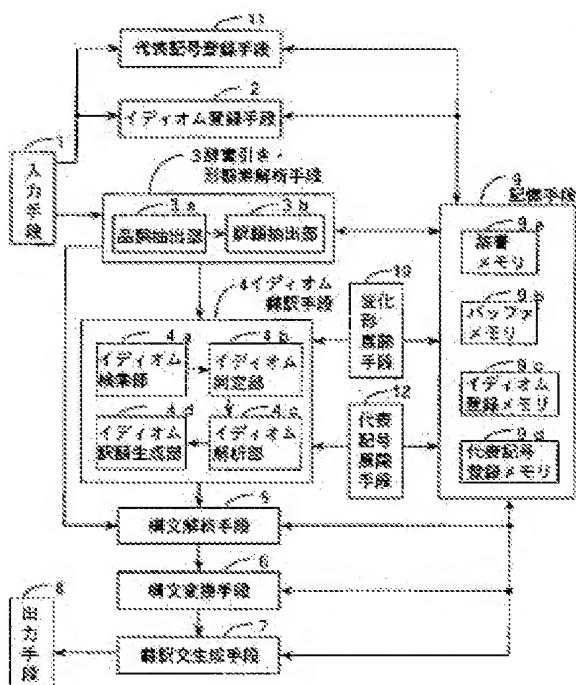
4-a イディオム検索部
 4-b イディオム固定部
 4-c イディオム解析部
 4-d イディオム記録生成部
 5 文字解析手段
 6 文字交換手段
 7 翻訳文生成手段
 8 出力手段
 9 記憶手段
 9-a 語書メモリ
 9-b バッファメモリ
 9-c イディオム登録メモリ

【図1】

18

9-d 代表記号登録メモリ
 10 変化形接頭手段
 11 代表記号登録手段
 12 代表記号隠匿手段
 3-1 メインCPU
 3-2 メインメモリ
 3-3 表示装置
 3-4 キーボード
 3-5 翻訳モジュール
 3-6 語書メモリ
 3-7 パス

【図5】

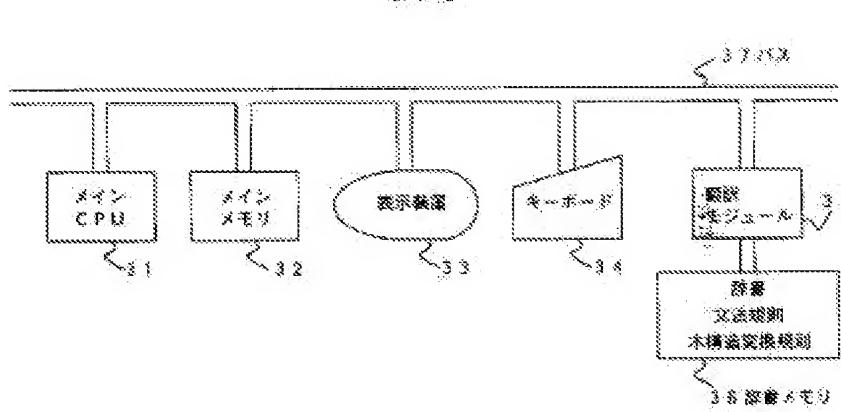


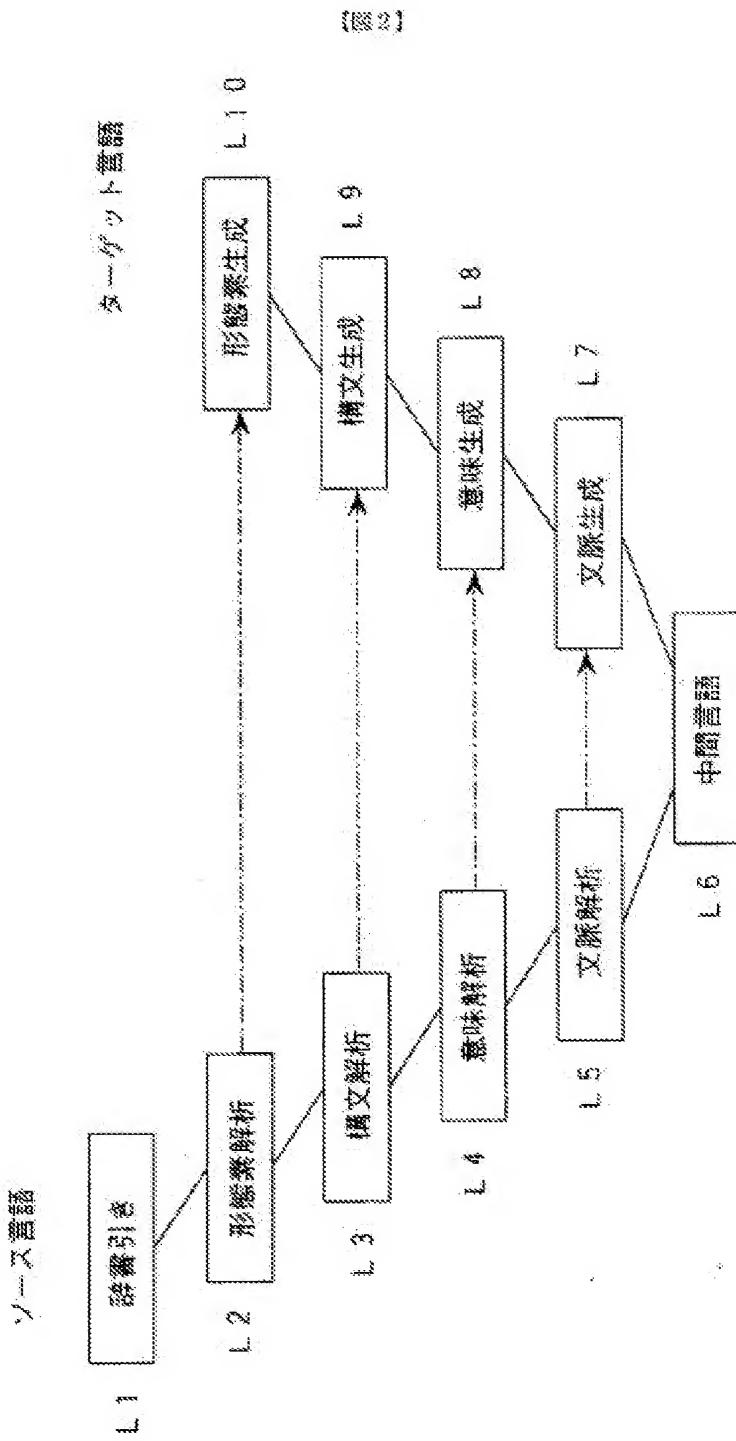
【図3】

例文：(The first time a person.)
 パッファA： 翻訳パッファ

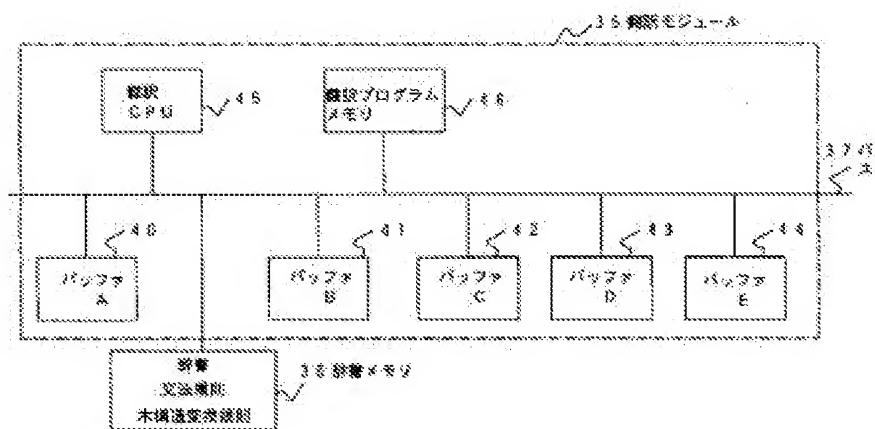
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851	852	853	854	855	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883	884	885	886	887	888	889	890	891	892	893	894	895	896	897	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912	913	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930	931	932	933	934	935	936	937	938	939	940	941	942	943	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960	961	962	963	964	965	966	967	968	969	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	988	989	990	991	992	993	994	995	996	997	998	999	1000

これはペンである。



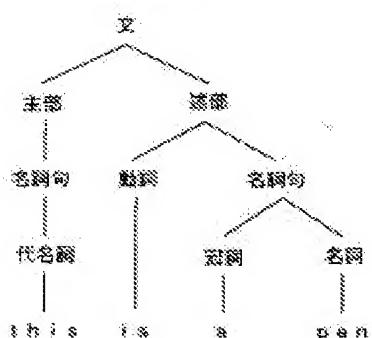


【図4】



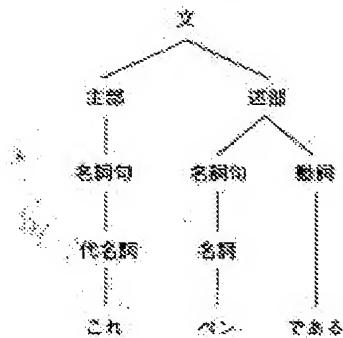
【図7】

バックファーム：構文解析結果のバックファーム



【図8】

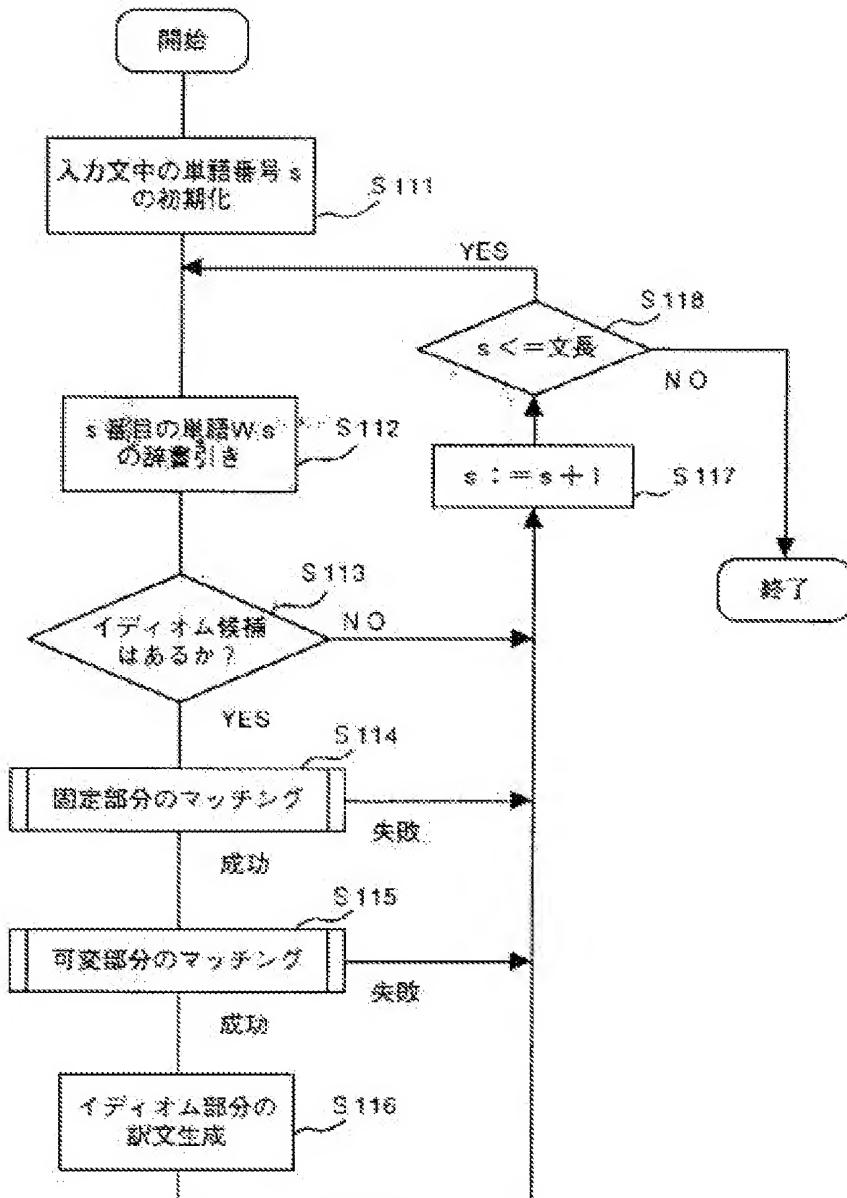
バックファーム：木変換した結果のバックファーム



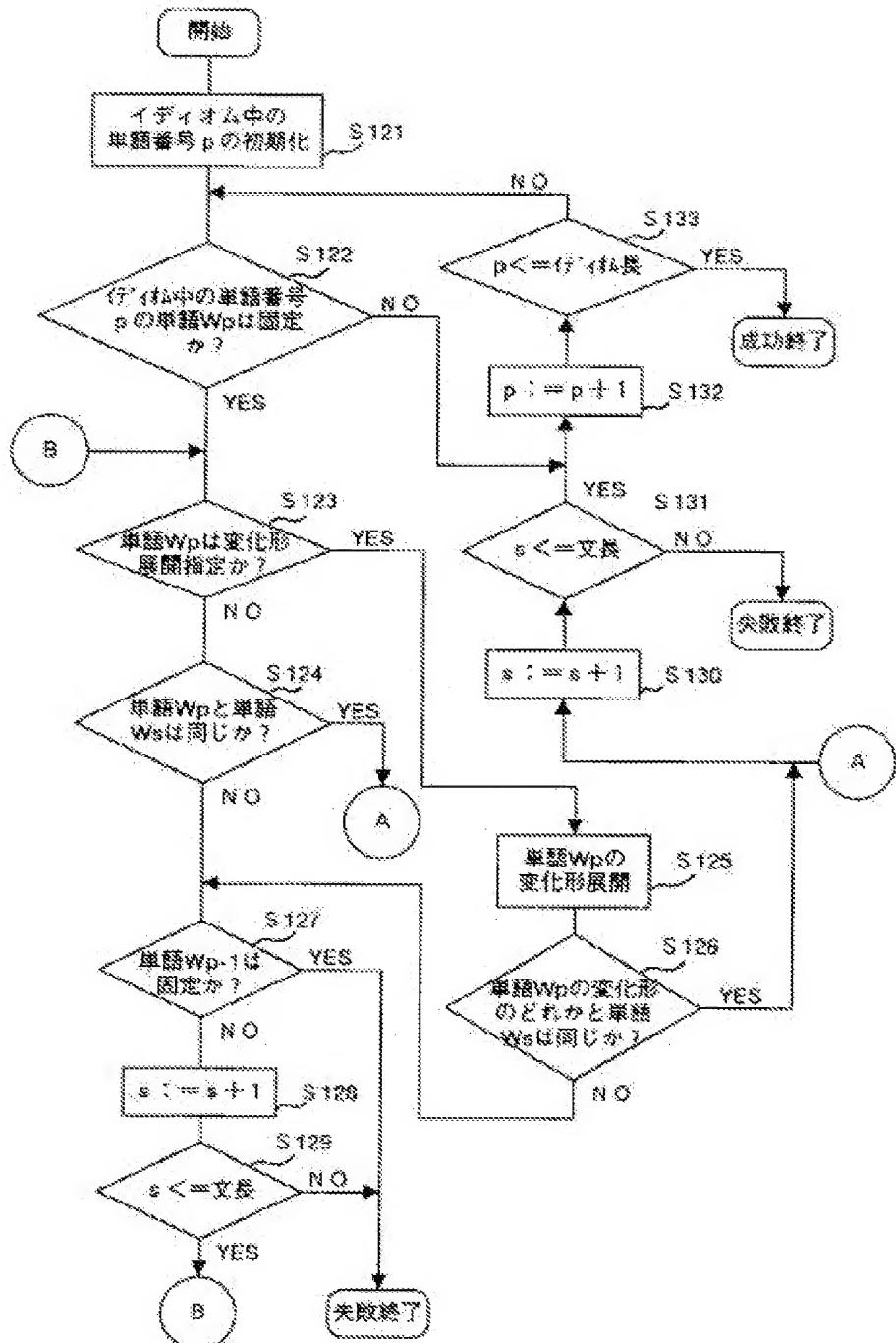
【図1.3】

原形	品詞	変化形
can	x	can, could
	v	can, cans, caned
	n	can, cans
be	v	is, was, were, will be, is not, was not, were not, will not be
hat	n	hat, hats

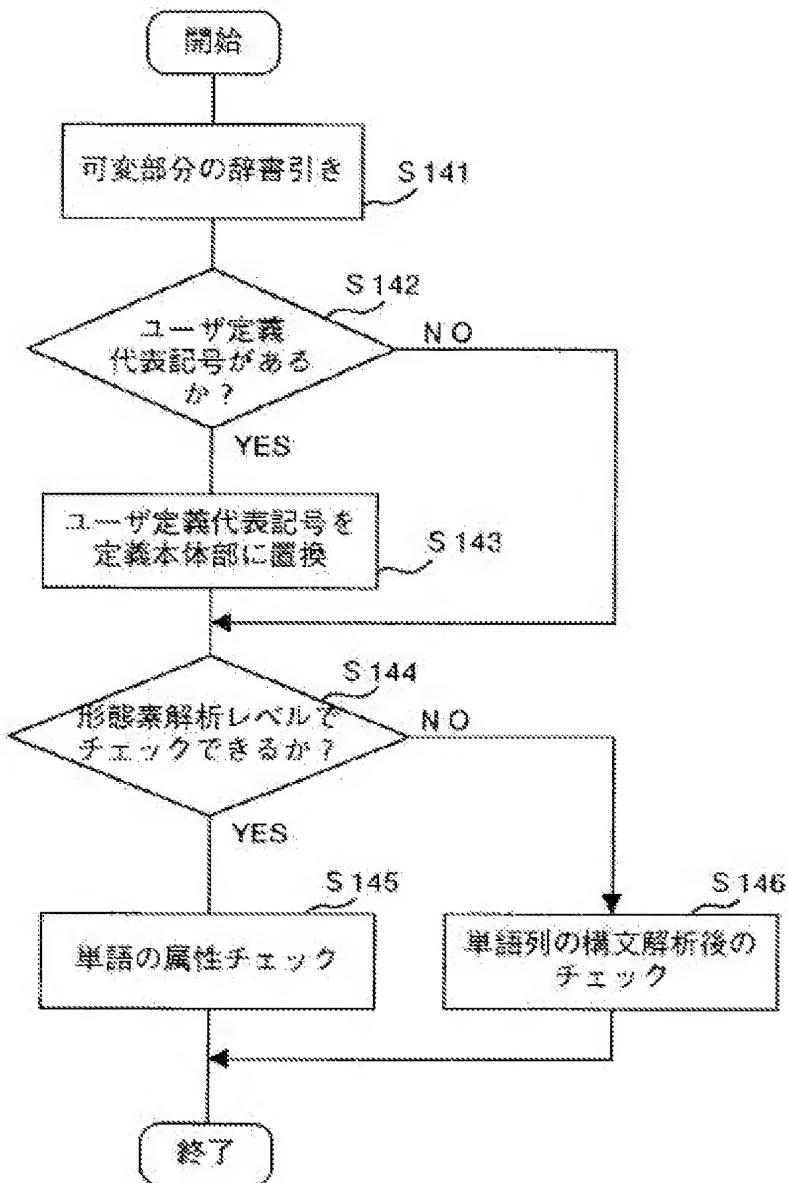
【図10】



[88] 64



【図1.2】



【図1-4】

属性	属性値	意味	チェック処理
品詞cat	n	名詞	形態素解析
	v	動詞	形態素解析
	s	形容詞	形態素解析
	q	副詞	形態素解析
	np	名詞句	構文解析
	vp	動詞句	構文解析
変化形inf	ing	進行形	形態素解析
	ed	過去形	形態素解析
	en	過去分詞形	形態素解析
	pl	複数	形態素解析
	sg	単数	形態素解析
	er	比較級	形態素解析
	est	最高級	形態素解析
意味sem	hum	人間	意味解析
	anim	可動物	意味解析
	prod	生産物	意味解析
	nat	自然	意味解析
	obj	機械的物体	意味解析
	place	物理的場所	意味解析
	time	時間	意味解析
用法use	deta	詳細をつくる	構文解析
	detals	詳細がつく	構文解析
	detail	詳細なし	構文解析
	~		

フロントページの続き

(72)発明者 九連見 繩
 大阪府大阪市阿倍野区長池町22番22号-シ
 ャープ株式会社内